METHODS, SYSTEMS, AND COMPUTER PROGRAM PRODUCTS FOR PROVIDING AUTOMATED CUSTOMER SERVICE VIA AN INTELLIGENT VIRTUAL AGENT THAT IS TRAINED USING CUSTOMER-AGENT CONVERSATIONS

CROSS-REFERENCE TO PROVISIONAL APPLICATION

This application claims the benefit of Provisional Application Serial No. 60/269,032, filed February 15, 2001, entitled *Methods, Systems, and Computer Program Products for Providing Automated Customer Service Via an Intelligent Virtual Agent that is Trained Using Customer-Agent Conversations*, the disclosure of which is hereby incorporated herein by reference in its entirety as if set forth fully herein.

BACKGROUND OF THE INVENTION

The present invention relates generally to the field of artificial intelligence and, more particularly, to the use of artificial intelligence to provide automated customer service.

As the Internet has evolved into a viable commercial medium, more businesses have developed Web sites to market and sell their products and/or services electronically. As businesses grow, call center and e-mail traffic may continue to escalate. This has resulted in customer service being a relatively fast growing segment of the customer relations management (CRM) market. It may, however, be expensive to provide numerous sales/customer service representatives to handle customer queries through call centers and/or e-mail systems. Hoping to deflect escalating call center and/or e-mail service expenses while enhancing the customer experience, businesses may turn to self-service solutions. Thus, there exists a need for

improved customer service systems, such as Web self-service solutions, that may allow businesses to increase operating efficiencies and turn customer service into a revenue-generating channel.

SUMMARY OF THE INVENTION

According to embodiments of the invention, a customer communication is responded to by receiving an utterance from the customer at an agent that executes on a data processing system. The agent uses a knowledge base that comprises information extracted from one or more exemplary conversations to generate a response to the received utterance. The agent then sends the generated response to the customer.

In other embodiments of the invention, the agent generates the response by analyzing the received utterance based on one or more prior utterances received from the customer, one or more prior responses sent from the agent to the customer, and/or the knowledge base.

In other embodiments of the invention, a conversation model is maintained that is representative of one or more prior utterances received from the customer and one or more prior responses that have been sent from the agent to the customer. The current state of the conversation model is updated based on the utterance received from the customer.

In other embodiments of the present invention, the utterance received from the customer and/or the conversation model may be sent to a customer service representative. The customer service representative may provide a notification of intent to communicate with the customer and may select one or more responses generated by the agent to send to the customer.

In still other embodiments of the present invention, the agent may receive a proposed response from the customer service representative and may determine whether the proposed response is appropriate to send to the customer. If the proposed response is determined to be inappropriate, then the proposed response may be sent to a supervisor for approval.

In further embodiments of the present invention, an agent may be trained to respond to a customer communication by compiling one or more exemplary conversations, which comprise an exchange of utterances. The conversations are

annotated to categorize the information contained therein and then processed using a machine learning engine to populate a knowledge base.

In still further embodiments of the present invention, a user is presented with multiple categories for annotating a conversation and, based on the user's input, parts of the conversations are associated with respective ones of the categories. The conversation parts may comprise sentences and words and it may be verified that all words that are determinative to the meaning of utterances comprising the conversation are annotated.

Although the present invention has been described above primarily with respect to method aspects of the invention, it will be understood that the present invention may be embodied as methods, systems, and computer program products.

BRIEF DESCRIPTION OF THE DRAWINGS

Other features of the present invention will be more readily understood from the following detailed description of specific embodiments thereof when read in conjunction with the accompanying drawings, in which:

FIG. 1 is a block diagram that illustrates communication networks for providing automated customer service in accordance with embodiments of the present invention;

FIG. 2 is a block diagram that illustrates data processing systems in accordance with embodiments of the present invention;

FIG. 3 is a software architecture block diagram that illustrate methods, systems, and computer program products for providing automated customer service in accordance with embodiments of the present invention; and

FIGS. 4 - 12 are flowcharts that illustrate exemplary operations of methods, systems, and computer program products for providing automated customer service in accordance with embodiments of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

While the invention is susceptible to various modifications and alternative forms, specific embodiments thereof are shown by way of example in the drawings and will herein be described in detail. It should be understood, however, that there is no intent to limit the invention to the particular forms disclosed, but on the contrary,

the invention is to cover all modifications, equivalents, and alternatives falling within the spirit and scope of the invention as defined by the claims. Like reference numbers signify like elements throughout the description of the figures.

The present invention may be embodied as methods, systems, and/or computer program products. Accordingly, the present invention may be embodied in hardware and/or in software (including firmware, resident software, micro-code, *etc.*). Furthermore, the present invention may take the form of a computer program product on a computer-usable or computer-readable storage medium having computer-usable or computer-readable program code embodied in the medium for use by or in connection with an instruction execution system. In the context of this document, a computer-usable or computer-readable medium may be any medium that can contain, store, communicate, propagate, or transport the program for use by or in connection with the instruction execution system, apparatus, or device.

The computer-usable or computer-readable medium may be, for example but not limited to, an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, device, or propagation medium. More specific examples (a nonexhaustive list) of the computer-readable medium would include the following: an electrical connection having one or more wires, a portable computer diskette, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), an optical fiber, and a portable compact disc read-only memory (CD-ROM). Note that the computer-usable or computer-readable medium could even be paper or another suitable medium upon which the program is printed, as the program can be electronically captured, via, for instance, optical scanning of the paper or other medium, then compiled, interpreted, or otherwise processed in a suitable manner, if necessary, and then stored in a computer memory.

FIG. 1 is a block diagram that illustrates communication networks for providing automated customer service in accordance with embodiments of the present invention. An exemplary communication system **22** network architecture comprises a server data processing system **24** that is coupled to one or more customer data processing systems **26** over a network **28**. The server data processing system **24** is further coupled to a customer service representative (CSR) data processing system **32** and a supervisor data processing system **34** over a network **36**. Optionally, the server

data processing system **24** may be coupled to an intelligent virtual agent (IVA) data processing system **38** over a network **42**.

The server data processing system **24** may be configured with computational, storage, and control program resources for providing automated customer service in accordance with embodiments of the present invention. The IVAs that facilitate automated customer service may be implemented on the server data processing system **24** or, alternatively, on the separate data processing system **38**, which communicates with the server data processing system **24** over the network **42**. Customers communicate with the server data processing system **24** via data processing systems **26** over the network **28**. The customer data processing system **26** may be, for example, a desktop computer, a Personal Communications System (PCS) terminal that may combine a cellular radiotelephone with data processing and data communications capabilities; a personal digital assistant (PDA) that can include a radiotelephone, pager, Internet/intranet access, Web browser, organizer, calendar and/or a global positioning system (GPS) receiver; and a conventional laptop/palmtop receiver, and/or an Internet appliance that includes a communication transceiver. The customer data processing system **26** may represent a class of devices sometimes referred to as "pervasive computing" devices.

The CSR data processing system **32** and the supervisor data processing system **34** may be implemented in like fashion to the customer data processing system **26**. A CSR may communicate with a customer, with the assistance of an IVA residing on either the server data processing system **24** and/or the IVA data processing system **38**, through the CSR data processing system **32**, the server data processing system **24**, and the networks **36** and **28**. An IVA may communicate with a customer service supervisor as necessary via the supervisory data processing system **34** and the network **36**.

The server data processing system **24** and/or the IVA data processing system **38** may be implemented as a single processor system, a multi-processor system, or even a network of stand-alone computer systems. The server data processing system **24** and/or the IVA data processing system **38** may communicate with a data storage repository **44** for storing, for example, log files of conversations between IVAs/CSRs and customers. These files may be edited by a CSR or another party responsible for

maintaining the IVA and associated automated customer service software on the server data processing system **24** and/or the IVA data processing system **38** for use in training the IVA subsystem.

The networks **28, 36**, and **42** may represent global networks, such as the Internet, or other networks accessible by the general public. The networks **28**, **36**, and **42** may also, however, represent wide area networks, local area networks, Intranets, or other private networks, which are not accessible by the general public. Furthermore, the networks **28**, **36**, and **42** may represent a combination of public and private networks or a virtual private network (VPN). In view of the foregoing, even though networks **28**, **36**, and **42** are illustrated in **FIG. 1** as separate networks, the networks **28**, **36**, and **42** may nevertheless be embodied as a single network.

Although **FIG. 1** illustrates an exemplary communication system **22** network architecture that may be used to provide automated customer service in accordance with embodiments of the present invention, it will be understood that the present invention is not limited to such a configuration but is intended to encompass any configuration capable of carrying out the operations described herein.

Referring now to **FIG. 2**, an exemplary data processing system **48** architecture is illustrated, which may be used in embodiments of the server data processing system **24**, the customer data processing system **26**, the CSR data processing system **32**, the supervisor data processing system **34**, and/or the IVA data processing system **38**, in accordance with the present invention. The data processing system **42** may include input device(s) **52**, such as a keyboard or keypad, a display **54**, and a memory **56** that communicate with a processor **58**. The data processing system **48** may further include a storage system **62**, a speaker **64**, and an input/output (I/O) data port(s) **66** that also communicate with the processor **58**. The storage system **62** may include removable and/or fixed media, such as floppy disks, ZIP drives, hard disks, or the like, as well as virtual storage, such as a RAMDISK. The I/O data port(s) **66** may be used to transfer information between the data processing system **48** and another computer system or a network (*e.g.*, the Internet). These components may be conventional components such as those used in many conventional computing devices, which may be configured to operate as described herein.

FIG. 3 illustrates a processor **72** and a memory **74** that may be used in embodiments of the server data processing system **24** for providing automated customer service in accordance with the present invention. The processor **72** communicates with the memory **74** via an address/data bus **76**. The processor **72** may be, for example, a commercially available or custom microprocessor. The memory **74** is representative of the overall hierarchy of memory devices containing the software and data used to provide automated customer service in accordance with embodiments of the present invention. The memory **74** may include, but is not limited to, the following types of devices: cache, ROM, PROM, EPROM, EEPROM, flash, SRAM, and DRAM.

As shown in **FIG. 3**, the memory **74** may contain up to nine or more major categories of software and/or data: the operating system **78**, the communication program module **82**, the agent registry program module **84**, the routing controller program module **86**, the machine learning program module **88**, the knowledge base **92**, the IVA program module **94**, the natural language program module **96**, and the conversation log **98**. The operating system **78** controls the operation of the computer system. In particular, the operating system **78** may manage the computer system's resources and may coordinate execution of programs by the processor **72**. The communication program **82** may be configured to communicate information between the server data processing system **24** and the other data processing systems in the communication network **22** of **FIG. 1** using suitable communication protocol(s). The communication program **82** may also comprise code for providing a graphical user interface (GUI) to customers to facilitate communication therewith.

The agent registry **84** may be configured to manage the IVAs as a system resource. That is, the agent registry assigns an IVA to a new customer and returns the IVA to an available pool once the conversation between the IVA and a customer is complete. The routing controller **86** may be configured to communicate with the agent registry so as to provide a CSR access to a particular customer under the supervision of the IVA associated with that customer.

The machine learning program **88** may be configured to populate the knowledge base **92** by processing exemplary conversations between a customer and an IVA. More specifically, the machine learning program **88** may comprise a

conversation annotator module **102** that may be used to annotate conversations to thereby categorize the information contained in the conversations. The machine learning program **88** may process these annotated conversations to populate the knowledge base **92**. In addition to the information derived from the exemplary conversations, the knowledge base **92** may include online word dictionaries, programmed rules and fact tables, and/or information databases. In accordance with embodiments of the present invention, the machine learning program **88** may facilitate the incorporation of knowledge through synonym sets, hand programming, machine learning, and/or database lookups. The "knowledge" may be organized in terms of "phrase sets," which are similar to conventional grammatical phrases, but are generally less restrictive. For example, a phrase set hierarchy may be used in which a highest level includes utterances that may be recognized, a middle level includes types and organized groups of words, such as dates, addresses, etc., and a lower level includes individual words that may be recognized and/or associated with each other, such as synonyms.

The IVA program **94** may be configured to provide a plurality of IVAs that communicate with customers over the network **28** as shown in **FIG. 1**. The IVA program **94** comprises a response engine module **104** that may be used to generate responses to communications received from customers based on the information contained in the knowledge base **92**. In more detail, the response engine module **104** cooperates with the natural language program **96** to analyze communications received from customers based on information contained in the knowledge base **92**. The natural language program **96** takes text produced by a customer or customer service representative and produces responses that are appropriate for the current conversational context. The context may include the current web page, the current state of active web applications that are executing, and the conversational history. The natural language program **96** comprises a natural language understander module **106** and a natural language generator module **108**, which are described in more detail hereafter.

The natural language understander module **106** is provided with the text and context information described above and computes a new set of dialogue states. The natural language understander module **106** comprises three modules, in accordance

with embodiments of the present invention: a text annotator module **109**, an utterance classifier module **110**, and a dialogue manager module **111**.

The text annotator module **109** is provided with the text and context information, which it uses to create an annotated utterance. The annotated utterance may comprise a hierarchical representation that includes information used for understanding an utterance. This hierarchical representation may include labels for individual words for part of speech, alternate spellings, as well as the extraction of application-specific information such as dates, times, and product names.

The utterance classifier module **110** is provided with the annotated utterance and determines the most probable utterance classification category to which the utterance belongs. This determination may be based on the features noted above in the annotated utterance, as well as the elements of the conversational context noted above. The classification may be computed by various classification algorithms, such as Bayesian classification, heuristic rule-based classification, and/or neural network classification techniques, in accordance with embodiments of the present invention.

The dialogue manager module **111** is provided with the classification computed by the utterance classifier module **110**, as well as the annotated utterance and the current context information. The dialogue manager **111** determines a set of probable dialogue states to which the conversation has transitioned. Its representation may be controlled by computational mechanisms, such as a finite state machine and/or a hidden Markov model.

The resulting set of dialogue states, with probabilistic annotations, as well as the annotated utterance and the other elements of the context information, are used by the natural language generator module **108** to create an utterance that is responsive to the user's utterance (*e.g.*, a question, request, and/or statement). The response utterance is then instantiated with context-specific information using techniques, such as template-based generation and/or grammar-based generation. The natural language generator module **108** may create a textual response, as well as a context-specific response that may include elements such as web content, relevant search links, and/or rich media (*e.g.*, animation and/or video).

The conversation log **98** may be used to store conversations between customers and IVAs. The machine learning program **88** may process these logged conversations as discussed above so as to improve the knowledge base **92**.

It will be understood that the machine learning program **88**, the knowledge base **92**, the IVA program **94**, the natural language program **96**, and sub-combinations thereof may be implemented on the server data processing system **24** and/or the IVA data processing system **38** in accordance with embodiments of the present invention. Moreover, the conversation log **98** may be stored on the data storage repository **44** where the CSRs may access the stored conversations through the network **36**.

Although **FIG. 3** illustrates an exemplary software architecture that may be used for providing automated customer service in accordance with embodiments of the present invention, it will be understood that the present invention is not limited to such a configuration but is intended to encompass any configuration capable of carrying out the operations described herein.

Computer program code for carrying out operations of the present invention may be written in an object-oriented programming language, such as Java, Smalltalk, or C++. Computer program code for carrying out operations of the present invention may also, however, be written in conventional procedural programming languages, such as the C programming language or compiled Basic (CBASIC). Furthermore, some modules or routines may be written in assembly language or even micro-code to enhance performance and/or memory usage.

The present invention is described hereinafter with reference to flowchart and/or block diagram illustrations of methods, systems, and computer program products in accordance with exemplary embodiments of the invention. It will be understood that each block of the flowchart and/or block diagram illustrations, and combinations of blocks in the flowchart and/or block diagram illustrations, may be implemented by computer program instructions and/or hardware operations. These computer program instructions may be provided to a processor of a general purpose computer, a special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the computer or other programmable data processing apparatus, create means for implementing the functions specified in the flowchart and/or block diagram block or blocks.

These computer program instructions may also be stored in a computer usable or computer-readable memory that may direct a computer or other programmable data processing apparatus to function in a particular manner, such that the instructions

stored in the computer usable or computer-readable memory produce an article of manufacture including instructions that implement the function specified in the flowchart and/or block diagram block or blocks.

The computer program instructions may also be loaded onto a computer or other programmable data processing apparatus to cause a series of operational steps to be performed on the computer or other programmable apparatus to produce a computer implemented process such that the instructions that execute on the computer or other programmable apparatus provide steps for implementing the functions specified in the flowchart and/or block diagram block or blocks.

With reference to the flowcharts of **FIGS. 4 - 12** and the architectural block diagrams of **FIGS. 1** and **3**, exemplary operations of methods, systems, and computer program products for providing automated customer service, in accordance with embodiments of the present invention, will be described hereafter.

Referring now to **FIG. 4**, operations begin with an IVA receiving an utterance from a customer (block **112**). As used herein, "utterance" means a communicative act, which includes typing text, following Web links, verbal communication, *etc.*, and "conversation" means an exchange of utterances. Note that an utterance could also be a failure to communicate affirmatively, such as a prolonged silence or idle period. After receiving the utterance from the customer, the agent (*e.g.*, the IVA program **94**) generates a response based on information contained in the knowledge base **92** (block **114**). If, however, the IVA program **94** cannot generate a response because the utterance from the customer cannot be interpreted in light of the information currently stored in the knowledge base **92**, then the IVA program **94** sends the received utterance to a CSR for generation of a response. That is, the IVA program **94** escalates the responsibility for constructing a response to the received customer utterance to a CSR. The response engine module **104** then sends the generated response to the customer (block **116**).

Exemplary operations for generating the response to the customer utterance are illustrated in **FIG. 5**. The IVA program **94** passes the customer utterance to the natural language program **96**, which processes the customer utterance using the natural language understander module **106** and the natural language generator module **108**. The natural language understander module **106** analyzes the customer utterance using one or more of the following criteria: the knowledge base **92**, prior utterances

received from the customer, and prior responses sent from the agent to the customer (block **122**). Advantageously, the prior utterances received from the customer and the prior responses sent from the agent to the customer may provide a contextual framework for analyzing the customer utterance. The natural language generator module **108** provides a response based on the results of the analysis to the response engine module **104**, which sends the response to the customer.

As part of the analysis performed by the natural language understander module **106**, the utterance received from the customer and/or part(s) of the utterance received from the customer are recognized based on information from the knowledge base **92**. For example, an utterance may be viewed as comprising a plurality of data strings and recognizing a part of the utterance may comprise recognizing one of the plurality of data strings and/or a sub-combination of the plurality of data strings. The natural language understander module **106** may also recognize an utterance and/or a part of an utterance by associating the utterance and/or the part of the utterance with an information type that corresponds to a predefined information arrangement and a predefined information meaning. Examples of information types may include dates, times, phone numbers, product model numbers and/or names, *etc.*

Referring now to **FIG. 6**, the IVA program **94** and the natural language program **96** cooperate to maintain a conversation model that has a current state that is representative of the conversation history between the agent and the customer (block **132**), in accordance with further embodiments of the present invention. The IVA program **94** and the natural language program **96** update the current state of the conversation model based on the utterance received from the customer and the response sent from the agent to the customer (block **134**). The conversation model may allow the IVA program **94** and/or the natural language program **96** to track the progress of the conversation between the agent and/or CSR and the customer and to induce structure to participate (*i.e.*, generate responses) in the middle of a conversation even if a CSR has controlled the conversation and generated responses up to this point in the conversation.

In accordance with further embodiments of the present invention, a CSR through the routing controller program **86** and the agent registry program **84** may monitor a conversation between an agent and a customer without taking control of the conversation. As illustrated in **FIG. 7**, the response engine module **104** sends to the

CSR the utterance received from the customer and the response to the received customer utterance sent from the agent to the customer (block **142**). The CSR may also observe the conversation state; therefore, the response engine module **104** sends the current state of the conversation module to the CSR (block **144**).

In addition to monitoring conversations, however, a CSR may take control of a conversation between a customer and an agent. Referring now to **FIG. 8**, the IVA program **94** receives a notification that the CSR intends to communicate with the customer (*i.e.*, take control of the conversation) (block **152**). When the CSR does not control the conversation (*e.g.*, when the CSR is in monitoring mode), the natural language generator module **108** may discard less favorable responses based on the analysis of the customer utterance (see **FIG. 5**) and may provide only the most favorable response to the response engine module **104** to be sent to the customer. When the CSR has seized control of the conversation, however, the natural language generator module **108** may generate more than one possible response to send to the customer (block **154**). The response engine module **104** sends these possible responses to the CSR (block **156**) to allow the CSR to select which response to send to the customer. The IVA program receives the CSR's selection of one of the possible responses (block **158**) and then the response engine module **104** sends the selected response to the customer (block **162**).

Referring now to **FIG. 9**, further embodiments of the present invention allow the CSR to propose a response to be sent to the customer. The IVA program **94** receives a proposed response from the CSR (block **172**) and then makes a determination if the proposed response is appropriate to send to the customer (block **174**). Thus, the agent may act as a filter to protect a customer from an inexperienced or angry CSR, for example. Specifically, the IVA program **94** sends the proposed response to a customer service supervisor for approval if the response is determined to be inappropriate (block **176**). If the IVA program **94** determines that the proposed response from the CSR is appropriate, then the response engine module **104** sends the proposed response to the customer (block **178**).

Referring now to **FIG. 10**, the IVA program **94** records utterances received from the customer and responses sent from the agent to the customer in the conversation log **98** (block **182**), in accordance with further embodiments of the

present invention. The machine learning program **88** may use these logged conversations to improve the knowledge base **92** as will be discussed in greater detail hereinafter. A CSR or other party responsible for maintaining the server data processing system **24** software may review conversations to determine if the agent sent an improper response to the customer (block **184**). If an improper response is detected, the reviewing party edits the conversation log to correct the improper response (block **186**) so that the machine learning program **88** may process the conversation containing the corrected response to incorporate the correct response into the knowledge base **92**.

Referring now to **FIG. 11**, exemplary operations for training an agent to respond to an utterance from a customer will be described hereafter. Operations begin by compiling one or more exemplary conversations (block **192**). These conversations may be edited conversations from the conversation log **98** as discussed hereinabove and/or they may be generated to illustrate common customer service exchanges between customers and customer service representatives. The conversation annotator module **102** annotates the compiled conversations to categorize the information contained therein (block **194**). Annotation of the conversations may be viewed as similarity based labeling of sentences, words, and/or phrases in the conversations. The machine learning program **88** processes the annotated conversations to extract information therefrom to populate the knowledge base **92** (block **196**). Optionally, the machine learning program **88** analyzes the conversations to verify that all determinative words in the conversation are annotated (block **198**). That is, words in the conversation that may be instrumental to a proper interpretation of the conversation may be required to have an annotation associated therewith.

FIG. 12 illustrates further embodiments of the present invention in which the conversation annotator module **102** presents a user (*e.g.*, a CSR or other party responsible for maintaining the server data processing system **24** software) with categories for annotating the conversations (block **202**). For example, the conversations comprise utterances and the utterances comprise sentences and words. Therefore, the user may be presented with a plurality of categories based on intent for annotating sentences and a plurality of categories based on semantic content for annotating words (block **204**).

The flowcharts of **FIGS. 4 - 12** illustrate the architecture, functionality, and operations of an exemplary implementation of the server data processing system **24** and/or the IVA data processing system **38** software. In this regard, each block represents a module, segment, or portion of code, which comprises one or more executable instructions for implementing the specified logical function(s). It should also be noted that in other embodiments, the functions noted in the blocks may occur out of the order noted in **FIGS. 4 - 12**. For example, two blocks shown in succession may be executed substantially concurrently or the blocks may be executed in the reverse order, depending on the functionality involved.

Many variations and modifications can be made to the preferred embodiments without substantially departing from the principles of the present invention. All such variations and modifications are intended to be included herein within the scope of the present invention, as set forth in the following claims.